Chain of Generation: Multi-Modal Gesture Synthesis via Cascaded Conditional Control

Zunnan Xu, Yachao Zhang †, Sicheng Yang, Ronghui Li, Xiu Li†

Tsinghua Shenzhen International Graduate School, Tsinghua University
University Town of Shenzhen, Nanshan District, Shenzhen, Guangdong, P.R. China
xzn23@mails.tsinghua.edu.cn, {yachaozhang, li.xiu}@sz.tsinghua.edu.cn

Abstract

This study aims to improve the generation of 3D gestures by utilizing multimodal information from human speech. Previous studies have focused on incorporating additional modalities to enhance the quality of generated gestures. However, these methods perform poorly when certain modalities are missing during inference. To address this problem, we suggest using speech-derived multimodal priors to improve gesture generation. We introduce a novel method that separates priors from speech and employs multimodal priors as constraints for generating gestures. Our approach utilizes a chain-like modeling method to generate facial blendshapes, body movements, and hand gestures sequentially. Specifically, we incorporate rhythm cues derived from facial deformation and stylization prior based on speech emotions, into the process of generating gestures. By incorporating multimodal priors, our method improves the quality of generated gestures and eliminate the need for expensive setup preparation during inference. Extensive experiments and user studies confirm that our proposed approach achieves state-of-the-art performance.

Introduction

Gesture synthesis is a significant area of research within the realm of human-computer interaction (HCI), with diverse applications across various fields such as movies, robotics, virtual reality, and digital humans (Kucherenko et al. 2021). It is a challenging task that requires accounting for the dynamic movements of the human body, as well as the underlying rhythm, emotion, and intentionality (Nyatsanga et al. 2023). In co-speech gesture generation, three key indicators have emerged: (i) generating gestures that synchronize with the audio and accurately depict the semantic content of the spoken text, (ii) generating gestures that are consistent with the speaker's style, and (iii) aligning the generated gestures with the speaker's intentions, including symbolic actions that may resemble sign language.

While substantial progress has been made in generating gestures synchronized to audio (Ginosar et al. 2019; Qian et al. 2021; Yazdian, Chen, and Lim 2022; Yang et al. 2023d; Ao, Zhang, and Liu 2023; Yang et al. 2023a), there has been limited exploration of emotive gesture generation that matches the rhythm and intention of the speech. Previous studies (Yoon et al. 2020; Liu et al. 2022a) have demonstrated the effectiveness of introducing more modalities for gesture synthesis. However, most of these studies have not fully explored the potential of multimodal gesture synthesis modeling. In CaMN (Liu et al. 2022a), facial blendshapes and emotion labels are incorporated as additional inputs and represented as embeddings that are concatenated with the existing inputs to generate more native and expressive gestures. However, as shown in Figure 1, when the number of input modalities decreases, the performance of the model experiences a notable decline. In practical applications, it is common to encounter scenarios where some modalities are missing partially. Retraining a new model to accommodate these missing modalities can be a costly endeavor.

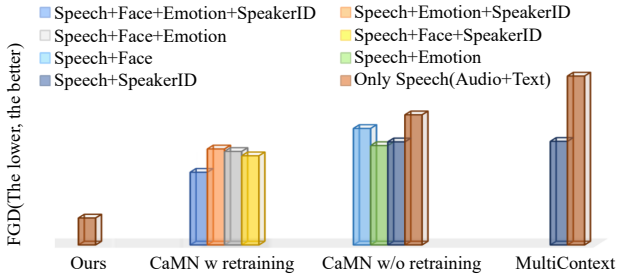

Figure 1: Performance comparison with limited modal during inference. The performance of existing multimodal methods is significantly hindered by the inadequate incorporation of multiple modalities during the inference stage. Our method addresses this limitation by utilizing prior information from speech to enable multimodal conditional control.

Recent studies (Yang et al. 2023b; Qi et al. 2023) have explored incorporating emotion labels into the generation of human gestures using various techniques (e.g., random mask, cross attention), resulting in diverse and emotive gestures. However, these approaches still rely on using additional modalities, such as emotion labels, during the inference process. The inconsistency between the assigned emotions and the context of speech may also lead to unnatural gestures. Furthermore, existing works overlook the importance of facial expressions in capturing speaker rhythm and

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
†Corresponding author

intent, leading to suboptimal gesture generation.

To address these problems, we propose a cascaded conditional control method for gesture synthesis. It aims to improve gesture generation by utilizing prior knowledge extracted from the speech, and maintain multimodal performance during inference when some modal data is missing (inference-efficient). Specifically, we introduce a novel method for extracting facial deformation and emotion-aware stylization priors from speech. By incorporating these priors, we greatly enhance the quality of gesture synthesis compared to previous methods. We incorporate an emotion-aware style injector to highlight the importance of emotional expression within the larger speech context. Firstly, we train a classifier to extract emotion information from speech. The extracted emotion information is then transformed into style features. We further introduce gesture adaptive layer normalization to apply the generated style features to all input features used by the decoder, resulting in more expressive gestures. Meanwhile, we suggest using a face decoder to convert speech into facial blendshapes. Considering that facial blendshapes provide important prior knowledge of facial deformations (e.g., lip language, the rate of deformation related to speech rhythm), we propose a temporal facial feature decoder and a rhythmic identification loss to extract facial deformation-aware priors from speech, which helps generate more rhythmic gestures. To ensure consistency in the generated results, we propose a chain-like generation framework, as shown in Figure 2, where the output of each stage serves as input for the next stage, ensuring a coherent generation across the face, body, and hands. Our approach eliminates the need for costly preparation during inference. This is achieved by training the model to generate facial blendshapes and emotion information from speech during the training phase. Furthermore, by integrating these generated modalities into the cascaded generation process, our method improves the quality of the resulting emotional gestures. The main contributions of our work are:

- We propose a novel inference-efficient approach that enables the model to learn to extract multimodal prior information during training, eliminating the need for costly setup preparation (e.g., blendshape capture devices) during inference. Simplification of modalities can greatly reduce the difficulty of application, as in most cases, it is not guaranteed that all modalities can be collected.
- We introduce a rhythmic identification loss to incorporate facial deformation as a guiding factor for generating gestures that synchronize with the speaker's speech rhythm.
- To enhance emotional expression in generated gestures, we propose an emotion-aware style injector. This component extracts emotional information from speech and incorporates it as a stylization prior to gesture synthesis.
- Extensive experiments and analyses demonstrate the effectiveness of the proposed approach.

Related Work

This work aims to develop a *multimodal conditional* approach for *co-speech gesture generation*. In this section, we summarize previous studies and discuss the relations and differences.

Multimodal Conditional Generation aims to generate content conditioned on various input modalities. This requires models to understand the relationships between different modalities and use them to guide the generation process (Qian et al. 2021). Mainstream approaches for achieving multimodal fusion include: (i) Attention mechanisms, which can model fine-grained inter-modality interactions (Li et al. 2021a; Zhang et al. 2022a; Li et al. 2023); (ii) Variational autoencoders, which can capture multimodal distributions (Liu et al. 2022b; Qi et al. 2023); (iii) Concatenation of representations from multiple modality-specific encoders (Yoon et al. 2020; Yang et al. 2022). Recent studies (Liu et al. 2022a; Yang et al. 2023b) have included emotional embeddings as inputs. Some pioneering works (Qi et al. 2023; Yin et al. 2023) achieve gesture diversity by learning emotion distributions and modifying emotion inputs during inference. However, these studies fail to consider the connection between emotion and the context in which speech occurs. Inconsistency between assigned emotions and the context of speech can result in unnatural gestures. Moreover, these methods rely on a large amount of modalities as input. They suffer a significant drop in performance when the number of input modalities decreases. Additionally, previous research has neglected the utilization of important priors that can be derived from facial blendshapes, leading to suboptimal gesture generation. Our approach to multimodal conditional generation focuses on extracting multimodal priors from speech. This allows the model to efficiently utilize information from the speech context and generate emotional gestures that align with the speech rhythm.

Co-speech Gesture Generation focuses on generating gestures based on speech input. Previous methods can be categorized into three types: (i) Linguistic rule-based methods convert speech into predefined gesture fragments and generate gestures using these rules (Cassell et al. 1994; Kopp and Wachsmuth 2004; Wagner, Malisz, and Kopp 2014); (ii) Statistical models that learn mapping rules from data and combine them with predefined gesture units to generate gestures (Kipp et al. 2007; Levine, Theobalt, and Koltun 2009; Levine et al. 2010); (iii) Deep learning methods that use neural networks to model the relationship between speech and gesture (Yoon et al. 2019; Yang et al. 2023c). While rule-based approaches can yield results that are easy to understand and control, they require a substantial amount of manual effort to create gesture datasets and engineer rules. Data-driven methods have become predominant for this task. Recent advances (Hu et al. 2022; Zhang et al. 2021) in deep learning have allowed neural networks to directly learn the complex relationships between speech and gestures from raw multimodal data (Zhang et al. 2022b). Previous research has shown that increasing the number of input modalities allows models to generate a wider range of expressive gestures (Yoon et al. 2020; Liu et al. 2022a). Although incorporating additional input modalities (e.g., facial blendshapes, emotion) improved performance, it also led to greater application difficulty during inference (e.g., requiring more capture devices, longer preprocessing times). Additionally, when there are multiple input modalities, there is often redundant modal information, allowing models to take short-

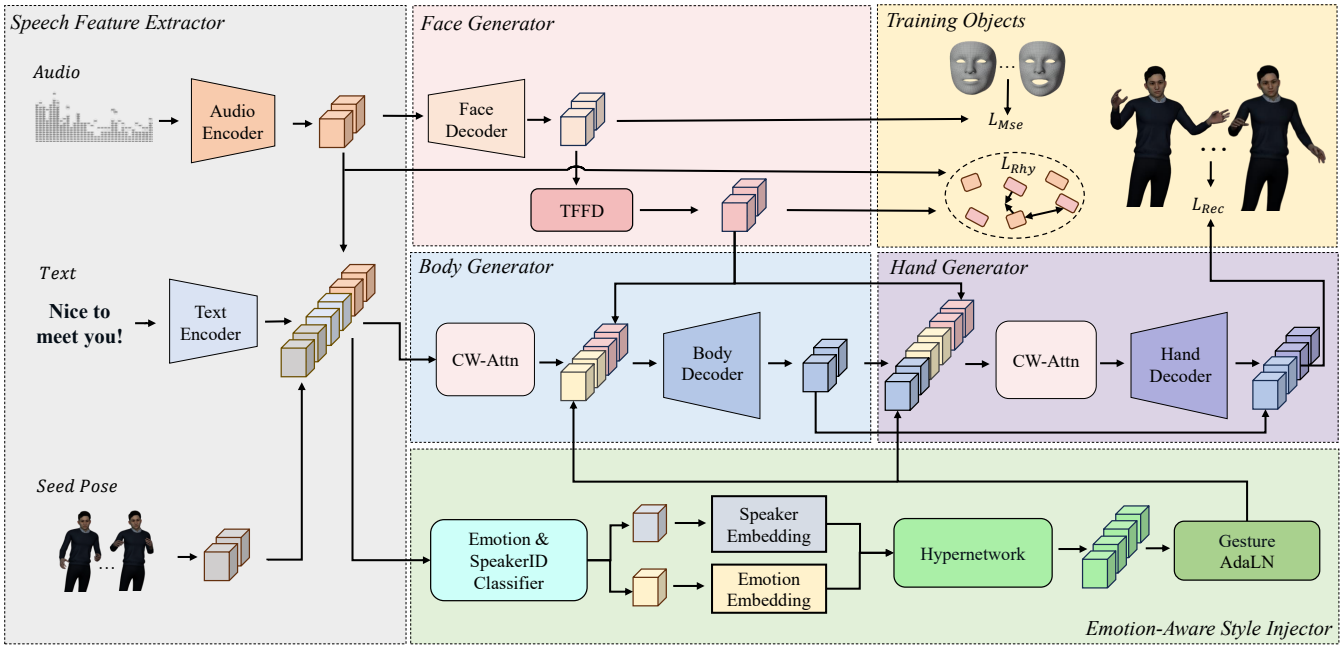

Figure 2: Overall architecture of the proposed CoG. Given audio sequences, text sequences, and seed poses, we use a chain-like framework to generate facial blendshapes, body movements, and hand gestures sequentially. We incorporate a classifier to extract emotion labels and speaker IDs from the speech, and use a hypernetwork to learn the style features. We further introduce gesture adaptive layer normalization to apply the generated style features to all input features used by the body and hand decoder. We leverage the ground truth of facial blendshapes as extra guidance during the training of the face generator to facilitate the transition from speech to facial blendshape. The features from the temporal facial feature decoder (TFFD) are utilized to calculate rhythmic identification losses, aiming to enhance the rhythm of generated gestures. (Best viewed in color.)

cuts without fully utilizing all modalities. To overcome these limitations, our method investigates the capacity of models to generate supplementary modalities exclusively from speech. This approach not only reduces inference costs but also enhances the utilization of multimodal features.

Methodology

Overall Framework

To reduce the dependence on multimodal information that must be complete and consistent with training data during model prediction (Lin et al. 2023), we aim to design a gesture synthesis method that can utilize multimodal prior information for better training, while still maintaining high inference performance when lacks partial modal information. Such a modal that eliminates the costly setup preparation during inference, can improve its flexibility and applicability. We propose a chain of generation method: Cascaded Gesture Synthesizer, which models the gesture in order from simple to complex, beginning with the facial blendshape, followed by the body, and concluding with the hand.

Specifically, as a common approach in speech-driven gesture synthesis, we use audio sequences $A = \{a_1, ..., a_N\}$, text sequences $T = \{t_1, ..., t_N\}$ and previous gestures as input (e.g. seed poses) to guide the continuous sequence of co-speech gestures denoted as $G = \{g_1, ..., g_N\}$, as inputs. Here, N represents the total number of frames, and $g_i \in R^{J\times 3}$ represents the 3D rotation pose state of the i-th frame. For speech input words, we use a pre-trained FastText (Bojanowski et al. 2017) to convert them into a word embedding set. Then, the word sets are fine-tuned by an encoder E_T to generate the text feature $T \in \mathbb{R}^{128}$. For audio feature extraction, we utilize a semi-supervised speech model, wav2vec2.0 (Baevski et al. 2020).

We found that it is difficult to maintain the performance during the inference phase if some modal information is missing. Therefore, we first introduce a linear projection layer at the end of the audio encoder to obtain the 128-dimensional latent audio feature $A \in \mathbb{R}^{128}$, for refining the audio feature. In order to guide the transition, we present a supervision signal (annotated facial blendshapes) during training to encourage the separation of facial blendshapes from speech. Then, we also extract rhythmic features from facial deformation using a temporal convolutional decoder and a rhythmic identification loss. Finally, to ensure accurate emotional expressions in the generated poses, we propose an emotion-aware style injector, which incorporates an additional classifier that generates emotion information from the speech input, and a hypernetwork to learn the style features. We further introduce gesture adaptive layer normalization to apply the generated style features to all input features used by the body and hand decoder. In this way, we utilize the deformation prior of facial blendshapes and the stylization

prior derived from speech to improve gesture synthesis.

Cascaded Gesture Synthesizer

Inspired by the concept of chain of thought and modal interaction (Wei et al. 2022; Xu et al. 2023), we propose to decompose the problem of speech-to-gesture generation into several stages. These stages include speech-to-emotion-based style recognition, speech-to-facial blendshape generation, facial blendshape-to-body generation, and body-to-hands generation. With emotion-based style features integrated into the latent features, our approach utilizes a chain-like modeling method, sequentially generating facial blendshapes, body movements, and hand gestures.

Face Generator. Our face generator is designed to generate a sequence of facial blendshapes $F = \{f_1, ..., f_N\}$ that are synchronized with audio features $A = \{a_1, ..., a_N\}$, where N denotes the total number of frames. To achieve this goal, we formulate 3D facial blendshape synthesis as a speech-driven sequence-to-sequence learning problem. Specifically, we adopt a 3-layer temporal convolutional decoder (FaceDec$(\cdot)$) to transfer audio features into facial blendshape F_t. Considering the temporal transformation of facial blendshape (e.g., the rate and degree of deformation), which can help the model understand the rhythm of speech, we decode the blendshapes into facial features $\hat{F}_t$ using a temporal facial feature decoder (TFFDec$(\cdot)$), which consists of 4 layers of temporal convolutional layers. The features will be further used to generate gestures in the body and hand generator. This allows us to capture local facial deformation over time, which can be formalized as:

$$\begin{aligned} F_t &= \text{FaceDec}(A_t), \\ \hat{F}_t &= \text{TFFDec}(F_t), \end{aligned} \tag{1}$$

where $t \in N$ denotes the t-th frame and N represents the total number of frames. Due to the temporal alignment of the annotated blendshape and speech, we add supervision on F_t to ensure the time consistency of the generated facial features with the speech.

Body and Hand Generator. As two generators have the same structure, we use body generator as an example to introduce this method for simplicity. In this generator, two modules (Channel-wise Attention Module and Body/Hand Decoer) are introduced detailed as follows.

Channel-wise Attention Module (CW-Attn). In order to facilitate the learning of features, we introduce the channel-wise attention module to perform adaptive reweighting of channel information to focus on more salient features. These attention weights selectively enhance informative channels via channel-wise multiplication of the latent features, which we refer to as "attended features". This module emphasizes useful channels while suppressing less informative ones. It contains average and max pooling layers to aggregate channel-wise statistics. The pooled outputs are fed into convolutional layers to generate a per-channel attention vector, activated by a sigmoid to range from 0 to 1, formulated as CW-Attn$(\cdot)$. We concatenate the multi-modal features and feed them into CW-Attn as:

$$C_t = \text{CW-Attn}([\hat{A}_t : \hat{T}_t : \hat{F}_t]), \tag{2}$$

where $t \in N$ denotes the i-th frame, N represents the total number of frames, and [:] denotes the concatenation operation. The attention vector is generated by convolving the pooled features, which selectively enhances salient channels via element-wise multiplication with the input.

Body Decoder. We adopt the separated, cascaded LSTM structure from previous work (Liu et al. 2022a) as decoder to capture the body/hand feature. Unlike these approaches, we incorporate multimodal priors as supplementary conditions to enhance gesture naturalness.

For reconstruction, the independent MLPs are added to the end of the body decoders to enable body synthesis. The process can be formalized as:

$$\begin{aligned} M_t &= \text{StyleInject}(C_t, S_t), \\ \hat{B}_t &= \text{BodyDec}(M_t), \\ B_t &= \text{BodyMlp}(\hat{B}_t). \end{aligned} \tag{3}$$

where StyleInject refers to the operation of injecting emotive and personal gesture style, which is given in (§).

Hand Decoder. The difference is that the input feature of the hand generator concatenates the output of the body generator, denoted as:

$$\hat{C}_t = \text{CW-Attn}([M_t : \hat{B}_t]). \tag{4}$$

We further model the hand gesture synthesis and get the $\hat{M}_t$, $\hat{H}_t$, and H_t same as body decoder.

Emotion-Aware Style Injector

We introduce an emotion-aware style injector to enable more expressive gestures. Contrary to previous works (Liu et al. 2022a; Yang et al. 2023b) that simply converts emotional labels into embedded features and concatenates them with other modal features as inputs. We consider emotional gesture generation as a stylized task and utilize a classifier to derive emotional information from speech.

Emotion & Speaker ID Classifier. To extract emotional information from the speech input, we first train a classifier on the training set using speech, emotion labels and speaker IDs. The classifier consists of a 3-layer temporal convolutional network with two linear projection layers, enabling the prediction of probabilities. We utilize cross-entropy loss to optimize the alignment between the predicted probability and the true emotion and speaker category. Then, we freeze the weights of the classifier and use its predictions as a guide to learn the style vector for the cascaded gesture synthesizer.

Gesture Adaptive Layer Normalization. We build upon the concept of adaptive layer normalization (Huang and Belongie 2017) and propose GestureAdaLN for stylized gesture generation. GestureAdaLN utilizes a hypernetwork to leverage inter-speaker and inter-emotion priors. The hypernetwork takes speaker and emotion embeddings as inputs and employs a 2-layer temporal convolution network to generate a style vector S_t for representing gesture style. Through two linear projection layers ($f(\cdot)$ and $g(\cdot)$), the style vector S_t is mapped to channel-wise mean and standard deviation parameters. These parameters are then used

to modulate the latent feature X:

$$\begin{aligned} S_t &= \text{HyNet}(E_t, I_t), \\ \hat{X} &= f(S_t) \cdot X + g(S_t), \end{aligned} \tag{5}$$

where $\text{HyNet}(\cdot)$ denotes the hypernetwork, and E_t and I_t represent the emotion labels and speaker ids at frame t. By considering the speaker identity and emotions, our method effectively captures the style differences in gesture content across different contexts, resulting in gestures that better align with the current speech content.

Training Objective

Rhythmic Identification Loss. Given the significance of a speaker's speech rhythm in their nonverbal communication, we suggest a novel method to aid in the generation of gestures from audio by integrating speech rhythms. A key insight is that the rhythm and prosody of speech provide important cues for generating natural gestures, while facial blendshapes provide additional contextual information about the tone and intent. Therefore, we apply the InfoNCE loss (Chen et al. 2020) to encourage temporal synchronization between the facial features $\hat{F}$ and audio features $\hat{A}$ to extract the speaking rhythm. Specifically, we compute the InfoNCE loss between encoded representations of the two modalities, which are obtained using neural network encoders f and g, respectively. This acts as a multimodal alignment loss that matches rhythmic cues in the audio with facial expressions, thereby generating better features with rhythmic information. The loss is defined as:

$$\ell_{\text{Rhy}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(f(\hat{F}_i), g(\hat{A}i))/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(f(\hat{F}_i), g(\hat{A}_j))/\tau\right)}, \tag{6}$$

where N is the number of frames in the aligned sequences, and $\hat{F}_i$ and $\hat{A}_i$ are the facial and audio features at the i-th frame, respectively. $\text{sim}(\cdot)$ denotes cosine similarity between the encoded representations, and τ represents the temperature hyperparameter. The synchronization loss encourages neural network encoders to learn representations that capture meaningful correlations between two modalities while disregarding irrelevant variations from the speech.

Face Reconstruction Loss. We utilize the mean squared error (MSE) loss as the face blendshape reconstruction loss. Specifically, the loss is defined as:

$$\ell_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{F}_i - \hat{\mathbf{F}}_i)^2, \tag{7}$$

where N indicates the number of frames, $\mathbf{F}_i$ denotes the i-th frame of ground truth blendshape parameters, and $\hat{\mathbf{F}}_i$ denotes the corresponding predicted blendshapes by our model. By minimizing the MSE during training, we aim to improve the accuracy and fidelity of the facial blendshapes in capturing facial expressions and deformations. This, in turn, enables the generation of expressive gestures by cascaded gesture synthesizer.

Body & Gesture Reconstruction Loss. For body and gesture generator, we utilize the L1 loss as the reconstruction loss function. The L1 loss measures the absolute difference between the predicted and ground truth values of the body and gesture parameters, providing a robust and efficient metric for reconstruction quality. The losses are defined as:

$$\begin{aligned} \ell_{\text{rec}}^{B} = \mathbb{E}\left[\left\|\mathbf{B} - \hat{\mathbf{B}}\right\|_1\right], \ell_{\text{rec}}^{H} = \mathbb{E}\left[\left\|\mathbf{H} - \hat{\mathbf{H}}\right\|_1\right], \\ \ell_{\text{Rec}} = \ell_{\text{rec}}^{B} + \alpha\ell_{\text{rec}}^{H}, \end{aligned} \tag{8}$$

where a weight α is adopted to balance the body and hands penalties, and $\mathbb{E}$ indicates the maximum likelihood estimation. Our objective is to improve the quality of the generated gesture by minimizing the L1 loss during training. This, in turn, enhances the model's capability to capture the dynamics and subtle nuances of human motion.

The Overall Objective. In summary, the overall optimization objective for our proposed method is formalized as:

$$\mathcal{L} = \lambda_{\text{Rhy}}\ell_{\text{Rhy}} + \lambda_{\text{Mse}}\ell_{\text{Mse}} + \lambda_{\text{Rec}}\ell_{\text{Rec}}. \tag{9}$$

where $\lambda_{\text{Rhy}} = 1$, $\lambda_{\text{MSE}} = 1000$ and $\lambda_{\text{Rec}} = 500$.

Experiments

Experiments Setting

Dataset. To evaluate the effectiveness of each component in our approach, we conducted comprehensive experiments on a large-scale multimodal dataset called BEAT (Body-Expression-Audio-Text) (Liu et al. 2022a). The dataset comprises 76 hours of multi-modal data captured from 30 speakers conversing in four different languages while expressing eight distinct emotions. This includes conversational gestures accompanied by facial expressions, emotions, and semantics, as well as annotations for audio, text, and speaker identity. To ensure a fair comparison, we followed CaMN (Liu et al. 2022a) and utilized approximately 16 hours of speech data from English speakers. Additionally, we followed the established practice of dividing the dataset into separate training, validation, and testing subsets, while maintaining the same data partitioning scheme as in previous work to ensure the fairness of the comparison.

Implementation Details. We use the Adam optimizer with an initial learning rate of 0.00025, and set the batch size to 512. To ensure a fair comparison, we use N = 34 frame clips with a stride of 10 during training. The initial four frames are used as seed poses, and the model is trained to generate the remaining 30 poses, which correspond to a duration of 2 seconds. Our models utilize 47 joints in the BEAT dataset, including 38 hand joints and 9 body joints. The latent dimensions of the facial blendshape, audio, text, and gesture features are all set to 128, while the speaker embedding and emotion embedding are set to 8. We set $\tau = 0.1$ in the rhythmic identification loss. All experiments are conducted using NVIDIA A100 GPUs. *More analysis results about hyperparameter are given in supplementary materials.*

Evaluation Metrics

Fréchet Gesture Distance (FGD) We used the Fréchet Gesture Distance (Yoon et al. 2020) to evaluate the distribution distance between the synthesized and ground truth gestures. To compute this metric, we utilize the autoencoder pre-trained by BEAT (Liu et al. 2022a).

Methods	FGD ↓	SRGR ↑	BeatAlign ↑
Seq2Seq	261.3	0.173	0.729
Speech2Gesture	256.7	0.092	0.751
MultiContext	176.2	0.195	0.776
Audio2Gesture	223.8	0.097	0.766
CaMN	123.7	0.239	0.783
TalkShow	91.00	-	0.840
GestureDiffuCLIP	85.17	-	-
CoG (ours)	**45.87**	**0.308**	**0.931**

Table 1: Comparison with methods (Yoon et al. 2019; Ginosar et al. 2019; Yoon et al. 2020; Li et al. 2021a; Liu et al. 2022a; Yi et al. 2022; Ao, Zhang, and Liu 2023) in the term of FGD, SRGR and BeatAlign. All methods are trained on BEAT datasets. ↓ denotes the lower the better while ↑ denotes the higher the better. The best results are in bold.

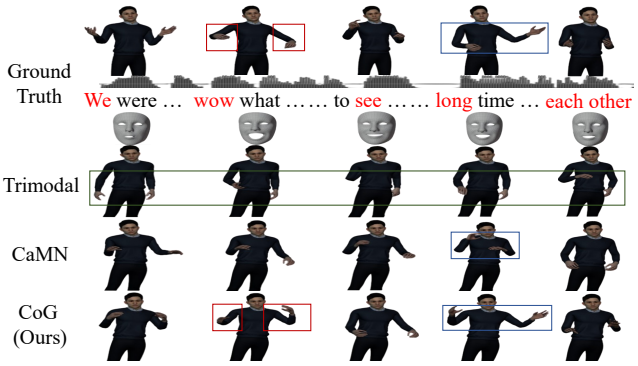

Figure 3: Visualization of our predicted 3D gestures against various baseline methods. The results of different methods are presented in separate rows, with each row representing the generated results of a method at different time frames.

Semantic-Relevant Gesture Recall (SRGR). We adopt the Semantic-Relevant Gesture Recall metric (Liu et al. 2022a) to evaluate the semantic relevance of generated gestures. SRGR leverages the semantic scores as weights for the Probability of Correct Keypoint metric between the generated gestures and the ground truth gestures. It can accurately capture the semantic aspects related to the generated gestures.
Beat Alignment Score (BeatAlign). To evaluate the correlation between gestures and audio, we utilized the Beat Alignment Score (Li et al. 2021b) to calculate the similarity between gesture beats and audio beats. BeatAlign provides a measure of alignment between the two modalities.

Qualitative Results

We compared our proposed method with existing multimodal gesture synthesis methods on the BEAT dataset in Table 1. Our approach achieves competitive performance across all metrics when compared to the state-of-the-art methods, which validates the effectiveness of our cascaded conditional framework for this task. In our method, we adopt an incremental modeling approach for human gestures. We start by generating facial blendshapes, then move on to the body, and finally the gestures. The chain-like generation pipeline enables lower FGD, leading to high-fidelity gesture reconstruction by gradually modeling gestures from simple to complex. Additionally, we effectively utilize multimodal information by decoupling stylization and rhythm cues from speech, resulting in significant improvements in the SRGR and BeatAlign scores.

Ablation Study

We validate the effectiveness of our proposed approach by conducting ablation studies on different components of our proposed method.
Effect of Cascaded Gesture Synthesizer. We evaluated the effectiveness of the cascaded gesture synthesizer through experiments conducted under various settings, as shown in Table 2. We examined the significance of the chain structure by conducting ablative experiments on the face generator. Specifically, "$-$ face$_{cog}$" refers to the method that does not join the face generator in our cascaded framework. In this setting, the facial blendshapes are converted into embeddings using a facial encoder and then concatenated with the encoding features of other modalities. "$+$ face$_{cog}$" represents the method that incorporates a face generator, where facial blendshape is included as a generation condition using the temporal facial feature decoder. The results of the ablation study show that incorporating the temporal facial feature decoder to process the generated features leads to a notable 31.7% decrease in FGD. We also examine the impact of integrating a channel-wise attention module, which highlights important channels while suppressing less informative ones. By incorporating the channel-wise attention module, indicated as "+ cw-attn," we achieve adaptive reweighting of channel information, giving higher priority to more prominent features and leading to enhanced overall metrics.
Effect of Emotion-Aware Style Injector. We validated the effectiveness of the emotion-aware style injector, as shown in Table 2, the method labeled as "$+$ style" includes the emotion-aware style injector to enable more expressive gestures. The ablation results demonstrate that incorporating emotional information from speech as a stylized prior enhances our method's capability to generate natural and contextually appropriate gestures. Moreover, the results confirm that considering gesture synthesis as a stylized task can enhance the expressiveness of the generated gestures. This improvement is clearly demonstrated in the overall metrics, particularly in the case of BeatAlign, which experienced a

Settings	FGD↓	SRGR↑	BeatAlign↑
$-$ face$_{cog}$	88.50	0.228	0.762
$+$ face$_{cog}$	60.44	0.240	0.834
$+$ cw-attn	58.74	0.238	0.842
$+$ style	52.36	0.241	0.916
$+ \mathcal{L}_{Rhy}$	45.87	0.308	0.931

Table 2: Ablation study on different components of our proposed method. ↓ denotes the lower the better, and ↑ denotes the higher the better.

significant increase of 8.79%.

Effect of Rhythmic Identification Loss. We validate the effectiveness of the rhythmic identification loss, as listed in Table 2, "$+\mathcal{L}_{Rhy}$" signifies the method that incorporates the rhythmic identification loss for further exploration of the rhythm of gesture. This improvement provides evidence for the correlation between facial blendshapes and gesture rhythm, as well as the successful separation of speech and facial features through contrastive learning. The benefits of investigating rhythm through facial features are clearly illustrated by the significant 27.8% increase in the SRGR score.

Qualitative Analysis

User Study. We conducted user study to evaluate the visual quality of the generated co-speech 3D gestures. For each compared method, we generated 10 results, and these gestures were converted into videos for evaluation by 22 participants. In each test, participants are presented with 20-second video clips synthesized by different models. Then, the participants are asked to provide ratings based on four dimensions: (i) naturalness, (ii) appropriateness, (iii) style correctness, and (iv) synchrony. In the naturalness test, participants are asked to evaluate the similarity of the generated gestures to those made by humans. This assessment primarily focuses on the naturalness and smoothness of the movements. In the appropriateness test, participants are required to assess the consistency of the gestures with the speech content, including both the literal content and the conveyed semantic meaning. In the style correctness test, participants are provided with emotion labels and asked to determine whether the generated gestures align with the intended style. In the synchrony test, participants assess the level of synchronization between gestures, speech rhythm, and accompanying audio and facial movements. They evaluate how well the gestures synchronize with these elements, ensuring a cohesive and harmonious overall presentation. We compared three methods, including MultiContext (Yoon et al. 2020), CaMN (Liu et al. 2022a), our method, and ground truth. As shown in Table 3, the average results demonstrate that our method achieves a significant advantage over the compared methods, performing better across all metrics.

Visualization. As illustrated in Figure 3, our method generates gestures that are more rhythmic and natural, aligning well with the speaking cadence. The gestures highlighted within the green rectangle indicate that previous methods were lacking in terms of diversity. In contrast, the gestures generated by our method exhibit a greater range of diversity. The gestures circled within the red rectangle indicate that our method is capable of recognizing gestures corresponding to expressive words. As shown in the figure, when the speaker utters an expressive word like "wow", the corresponding audio exhibits rhythmic fluctuations. Our method responds by performing a simultaneous upward movement of both hands, resembling the human gestures represented by the ground truth. In contrast, previous methods have been deficient in capturing this aspect. The gestures highlighted within the blue rectangle reveal that our method is capable of producing gestures that correspond to the intended meaning of the speaker. It can be observed that when the speaker mentions the word "long", our method generates a gesture of both hands spreading outwards, which aligns with the semantic meaning of the word. This gesture closely resembles the ground truth. In contrast, other methods lack a similar response in this regard. Our approach enables the generation of expressive gestures that not only appear natural but also synchronize with the rhythm of speech.

Methods	N↑	M↑	C↑	S↑
MultiContext	3.127	2.901	3.129	3.110
CaMN	3.588	3.261	3.225	3.414
Ours	3.916	3.624	3.541	3.685
Ground Truth	4.492	4.570	4.382	4.413

Table 3: The user study on naturalness (N, human likeness), matchness (M, the degree of consistency with the speech content), style correctness (C, with emotion labels), and synchrony (S, the level of synchronization with the speech rhythm). The rating score range is 1-5, with 5 being the best. ↑ indicates the higher the better.

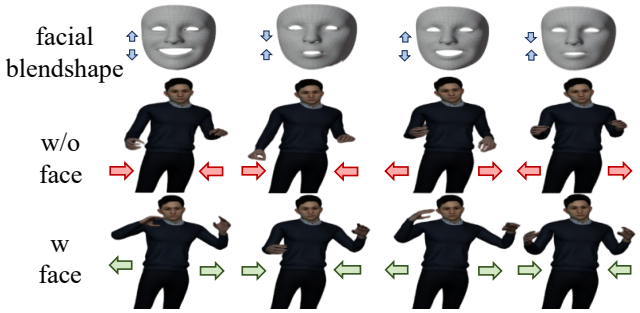

Figure 4: Visualization of the gestures generated by our method, the results of without/with a facial generator and rhythmic identification loss (denoted as *w/o face* and *w face*). The facial blendshape results are shown in the first row, and *w/o face* and *w face* are given in the second and third rows, respectively.

We compared the results generated without the face generator to those generated with the face generator in Figure 4. It can be observed that incorporating facial features aligns gesture rhythms with speech pace, demonstrating the advantages gained from including facial deformation priors.

Conclusion

In this study, we propose a framework to enhance the generation of 3D gestures by leveraging multimodal information from human speech. Our approach incorporates multimodal priors as constraints to enhance gesture generation. We adopt a chain-like modeling approach to sequentially generate facial blendshapes, body movements, and hand gestures. By incorporating rhythm cues from facial blendshapes and stylization priors into the generation process, our approach improves the quality of the generated gestures and reduces the number of modalities needed during inference.

Acknowledgements

This research was partly supported by Shenzhen Key Laboratory of next generation interactive media innovative technology (Grant No: ZDSYS20210623092001004), the China Postdoctoral Science Foundation (No.2023M731957), the National Natural Science Foundation of China under Grant 62306165.

References

Ao, T.; Zhang, Z.; and Liu, L. 2023. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *arXiv preprint arXiv:2303.14613*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146.

Cassell, J.; Pelachaud, C.; Badler, N.; Steedman, M.; Achorn, B.; Becket, T.; Douville, B.; Prevost, S.; and Stone, M. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, 413–420.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Ginosar, S.; Bar, A.; Kohavi, G.; Chan, C.; Owens, A.; and Malik, J. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3497–3506.

Hu, R.; Monebhurrun, V.; Himeno, R.; Yokota, H.; and Costen, F. 2022. An uncertainty analysis on finite difference time-domain computations with artificial neural networks: improving accuracy while maintaining low computational costs. *IEEE Antennas and Propagation Magazine*, 65(1): 60–70.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Kipp, M.; Neff, M.; Kipp, K. H.; and Albrecht, I. 2007. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7*, 15–28. Springer.

Kopp, S.; and Wachsmuth, I. 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds*, 15(1): 39–52.

Kucherenko, T.; Jonell, P.; Yoon, Y.; Wolfert, P.; and Henter, G. E. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *26th international conference on intelligent user interfaces*, 11–21.

Levine, S.; Krähenbühl, P.; Thrun, S.; and Koltun, V. 2010. Gesture controllers. In *ACM SIGGRAPH 2010 papers*, 1–11.

Levine, S.; Theobalt, C.; and Koltun, V. 2009. Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia 2009 papers*, 1–10.

Li, J.; Kang, D.; Pei, W.; Zhe, X.; Zhang, Y.; He, Z.; and Bao, L. 2021a. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11293–11302.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021b. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.

Li, R.; Zhao, J.; Zhang, Y.; Su, M.; Ren, Z.; Zhang, H.; Tang, Y.; and Li, X. 2023. FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10234–10243.

Lin, Y.; Han, H.; Gong, C.; Xu, Z.; Zhang, Y.; and Li, X. 2023. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*.

Liu, H.; Zhu, Z.; Iwamoto, N.; Peng, Y.; Li, Z.; Zhou, Y.; Bozkurt, E.; and Zheng, B. 2022a. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, 612–630. Springer.

Liu, X.; Wu, Q.; Zhou, H.; Du, Y.; Wu, W.; Lin, D.; and Liu, Z. 2022b. Audio-Driven Co-Speech Gesture Video Generation. *arXiv preprint arXiv:2212.02350*.

Nyatsanga, S.; Kucherenko, T.; Ahuja, C.; Henter, G. E.; and Neff, M. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*, volume 42, 569–596. Wiley Online Library.

Qi, X.; Liu, C.; Li, L.; Hou, J.; Xin, H.; and Yu, X. 2023. EmotionGesture: Audio-Driven Diverse Emotional Co-Speech 3D Gesture Generation. *arXiv preprint arXiv:2305.18891*.

Qian, S.; Tu, Z.; Zhi, Y.; Liu, W.; and Gao, S. 2021. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11077–11086.

Wagner, P.; Malisz, Z.; and Kopp, S. 2014. Gesture and speech in interaction: An overview.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Xu, Z.; Chen, Z.; Zhang, Y.; Song, Y.; Wan, X.; and Li, G. 2023. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17503–17512.

Yang, S.; Wang, Z.; Wu, Z.; Li, M.; Zhang, Z.; Huang, Q.; Hao, L.; Xu, S.; Wu, X.; Yang, C.; et al. 2023a. UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1033–1044.

Yang, S.; Wu, Z.; Li, M.; Zhang, Z.; Hao, L.; Bao, W.; Cheng, M.; and Xiao, L. 2023b. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. *arXiv preprint arXiv:2305.04919*.

Yang, S.; Wu, Z.; Li, M.; Zhang, Z.; Hao, L.; Bao, W.; and Zhuang, H. 2023c. QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speech-Driven Gesture Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2321–2330. IEEE.

Yang, S.; Wu, Z.; Li, M.; Zhao, M.; Lin, J.; Chen, L.; and Bao, W. 2022. The ReprGesture entry to the GENEA Challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 758–763.

Yang, S.; Xue, H.; Zhang, Z.; Li, M.; Wu, Z.; Wu, X.; Xu, S.; and Dai, Z. 2023d. The DiffuseStyleGesture+ entry to the GENEA Challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, 779–785.

Yazdian, P. J.; Chen, M.; and Lim, A. 2022. Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3100–3107. IEEE.

Yi, H.; Liang, H.; Liu, Y.; Cao, Q.; Wen, Y.; Bolkart, T.; Tao, D.; and Black, M. J. 2022. Generating Holistic 3D Human Motion from Speech. *arXiv preprint arXiv:2212.04420*.

Yin, L.; Wang, Y.; He, T.; Liu, J.; Zhao, W.; Li, B.; Jin, X.; and Lin, J. 2023. EMoG: Synthesizing Emotive Co-speech 3D Gesture with Diffusion Model. *arXiv preprint arXiv:2306.11496*.

Yoon, Y.; Cha, B.; Lee, J.-H.; Jang, M.; Lee, J.; Kim, J.; and Lee, G. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6): 1–16.

Yoon, Y.; Ko, W.-R.; Jang, M.; Lee, J.; Kim, J.; and Lee, G. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, 4303–4309. IEEE.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022a. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; and Li, C. 2021. Perturbed Self-Distillation: Weakly supervised large-scale point cloud semantic segmentation. In *ICCV*, 15520–15528.

Zhang, Y.; Xie, Y.; Li, C.; Wu, Z.; and Qu, Y. 2022b. Learning All-In Collaborative Multiview Binary Representation for Clustering. *IEEE Transactions on Neural Networks and Learning Systems*.